Regresión de datos misteriosos

C Astudillo

*Universidad de Talca, Facultad de Ingeniería, Km 1 camino a los Niches, Curicó, Chile.
Email: castudillo@utalca.cl

Abstract

Se presenta un conjunto de datos cuyo origen se mantiene confidencial. Este conjunto consta de 481 columnas de números reales distribuidos en 93 instancias, siendo la columna designada como "Y" la variable objetivo a predecir.

El propósito de esta investigación es identificar y evaluar un conjunto de modelos de regresión con el fin de lograr predicciones más precisas para los datos proporcionados.

1 Problema Específico

Los datos proporcionados tienen un origen no revelado, pero corresponden a una aplicación real en el campo de la ingeniería. El problema consiste en encontrar un modelo de regresión para predecir de manera efectiva la variable dependiente (Y).

2 Hipótesis

Usted debe definir su hipótesis. Esta corresponde a una frase que diga que es lo que esperamos encontrar. La hipótesis implícitamente debe especificar como se van a efectuar las mediciones para determinar si la hipótesis es correcta o no. Dado que el origen de los datos es una incógnita, la hipótesis debe ser genérica. El objetivo del resto del estudio es validar o refutar dicha hipótesis.

3 Datos

Los datos para este estudio se encuentran públicamente disponibles de la siguiente URL: `https://goo.gl/64CUV7`

4 Pre-procesamiento de los Datos

Usted debe detallar las técnicas que se usan en la etapa de pre-procesamiento de los datos. Algunos de los métodos son la detección de datos faltantes, detección de datos atípicos, imputación de datos, normalización preliminar, etc.

5 Detección de datos atípicos

Quizás, para mejorar el desempeño de las predicciones, puede ser útil descartar algunas instancias que se encuentran fuera de la norma. Si es así, debe determinar el procedimiento que utilizó (por ejemplo boxplots), pero debe considerar que no debería eliminar como máximo un 5% de los datos.

6 Normalización en etapa de post-procesamiento

Una vez que se hayan eliminado los datos atípicos, los rangos de las variables pueden variar. Por esta razón se sugiere volver a realizar una normalización de los datos después de remover los datos atípicos.

7 Visualización

Para tener una mejor idea de cual es el patrón de los datos, usted debe presentar algún tipo de visualización de los datos. Algunos de los métodos de visualización mas populares incluyen los siguientes:

- Principal Component Analysis (PCA),
- Multi Dimensional Scaling (MDS)
- Self-Organizing Maps (SOM)
- t-Distributed Stochastic Neighbor Embedding (t-SNE)

8 Reducción de la Dimensionalidad

Una alternativa para mejorar la predicción es reducir el numero de dimensiones. En esta sección usted debe especificar que técnicas se usaron, como por ejemplo eliminación por correlaciones, reducción de dimensiones a través del algoritmo Boruta, PCA, etc.

9 Regresión

Aquí debe detallar las técnicas de regresión que utilizará para predecir los datos Y. Como mínimo debe utilizar al menos dos algormitos de regresión. Utilice algunas de las técnicas vistas en clases e investigue acerca algunas técnica que pudiese ser mas efectiva.

10 Medidas de desempeño

En esta etapa, usted debe explicar brevemente que medida de desempeño se usará. Algunas alternativas son:

r-cuadrado, Standard error (SE), Mean Squared Error (MSE), Mean Squared Error (RMSE), Etc.

En cada caso es necesario especificar la ecuación respectiva e idealmente una referencia.

Como consejo, es buena idea mostrar varias metricas, ya que un unica metrica generalmente no explica todo el comportamiento de un modelo. Sin embargo, generalmente, para efectuar una comparacion con otros modelos se debe escoger la metrica más significativa.

11 Validación

Especificar el tipo de validación a usar. Una Alternativa muy común, es utilizar el modelo de validación cruzada de 10 subconjuntos (10-fold CV).

12 Resultados

En esta sección usted debe detallar los resultados de los experimentos realizados. Esto incluye Boxplot para las validaciones, tabla con datos valores promedio y desviación estándar (entre parentesis) para cada uno de los modelos de predicción escogidos.

Frecuentemente, cuando se hace una comparación mas rigurosa, se efectúa un test de hipotesis para la metrica (por ejemplo los valores promedio) y asi se elige un modelo "ganador". Esto se hace comunicando un valor de significancia p-value que fortalece en gran medida los resultados. Sin embargo este paso queda como opcional.

13 Conclusiones

En esta sección usted debe resumir que es lo que se ha hecho y dar una respuesta a la hipótesis planteada. Además usted puede agregar un par de párrafos con comentarios finales que incluyen su opinión respecto a distintos aspectos del trabajo realizado.

14 Entrega y plazos

La tarea debe entregarse a mas tardar el día viernes 29 de diciembre en un documento PDF de no más de 4 páginas. El documento debe incluir un enlace al código fuente (Jupiter Notebook, Gitub, Google Colab son recomendados) que se encuentra en una URL **pública**. El código debe reproducir íntegramente lo realizado en el proyecto.

El documento PDF debe enviarlo al aula virtual.

15 Recursos utiles

Principal Component Analysis for analyzing the Iris dataset [Github] `https://github.com/castudil/exploratory-data-analysis/blob/main/S08%20PCA/S8-pca-iris.ipynb`github